A34274 PCT USA - 072854.0119

PATENT

#3

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Piekarski et al.
Serial No.	:	09/869,086
Filed	:	June 20 2001
For	:	DISTRIBUTED HIERARCHICAL SCHEDULING AND ARBITRATION FOR BANDWIDTH ALLOCATION

CLAIM FOR PRIORITY UNDER 35 U.S.C. § 119

I hereby certify that this paper is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Assistant Commissioner for Patents, Washington, D.C. 20231, on:

August 6, 2001
Date of Deposit

Ronald B. Hildreth — 19,498
Attorney Name — PTO Reg. No.

Signature — August 6, 2001
Date of Signature

Assistant Commissioner for Patents

Washington, D.C. 20231

Sir:

A claim for priority is hereby made under the provisions of 35 U.S.C. § 119 for the above-identified PCT application based upon Great Britain application 9828143.9 filed December 22, 1998, and International Application PCT/GB99/04007 filed December 1, 1999.

Respectfully submitted,

Ronald B. Hildreth
Patent Office Reg. No. 19,498

(212) 408-2544
Attorney for Applicants

Baker Botts L.L.P.
30 Rockefeller Plaza
New York NY 10112

The Patent Office

INVESTOR IN PEOPLE

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

GB99/4007

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 16 DEC 1999

Patents Form 1/77

Act 1977

THE PATENT OFFICE
D
22 DEC 1998
RECEIVED BY POST

The Patent Office

22DEC98 E413493-6 D00107
P01/7700 0.00 - 9828143.9

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference

M98/0661/GB

2. Patent application number
(The Patent Office will fill in this part)

22 DEC 1998

9828143.9

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

Power X Limited

Stafford Court
145 Washway Road
Sale
Cheshire M33 7PE

Patents ADP number *(if you know it)*

Great Britain

If the applicant is a corporate body, give the country/state of its incorporation

6803233002

4. Title of the invention

Distributed Hierarchical Scheduling and Arbitration for Bandwidth Allocation

5. Name of your agent *(if you have one)*

McNeight & Lawrence

"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

Regent House
Heaton Lane
Stockport
Cheshire SK4 1BS

Patents ADP number *(if you know it)*

0001115001

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*

a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body.
See note (d))

Yes

following items you are filing with this form.
Do not count copies of the same document

Continuation sheets of this form	
Description	6
Claim(s)	–
Abstract	–
Drawing(s)	3 x 2

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature McNeight & Lawrence. Date 21/12/1998

12. Name and daytime telephone number of person to contact in the United Kingdom

0161 480 6394

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.

b) Write your answers in capital letters using black ink or you may type them.

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

Distributed Hierarchical Scheduling and Arbitration for Bandwidth Allocation

continual growth of demand for manageable bandwidth in networks requires the development of new techniques in switch design which decouple the complexity of control from the scale of the port count and aggregate bandwidth. This paper describes a switch architecture and a set of methods which provide the means by which switches of arbitrary size can be constructed while maintaining the ability to allocate guaranteed bandwidth to each possible connection through the switch.

A digital switch is used to route data streams from a set of source components to a set of destination components. A cell based switch operates on data which is packetised into streams of equal size cells. In a large switch, the routing functions can be implemented hierarchically: sets of lower bandwidth ports are aggregated into a smaller number of higher bandwidth ports which are then interconnected in a central switch. This document also describes how the bandwidth allocation technique can be applied in such a hierarchical switch.

A scheduling and arbitration process for use in a digital data switching arrangement of the type in which a central switch under the direction of a master control provides the cross-connection between a number of high-bandwidth ports to which are connected on the ingress side of the central switch ingress multiplexers, one for each high-bandwidth input port and on the egress side of the central switch one egress demultiplexer for each high-bandwidth output port, each ingress multiplexer includes a set of N input queues serving N low-bandwidth data sources and a set of M virtual output queues one for each low-bandwidth output data source *characterised in that* the scheduling and arbitration arrangement includes three bandwidth allocation tables, the first of which the, ingress port table (IPT), is associated with the input queues having NxM entries each arranged to define bandwidth allocation for a particular virtual output queue, the second of which, the egress port table (EPT), is associated with the virtual output queues having M entries each arranged to define the bandwidth allocation of a high-bandwidth port of the central switch to a virtual output queue and a third of which, the central allocation table (CAT) located in the master control having $(M/N)^2$ entries each of which specifies the weights allocated to each possible connection through the central switch.

According to a feature of the invention there is provided a scheduling and arbitration process in which the scheduling of the input queues is performed in accordance with an N way weighted round robin.

According to a further feature of the invention there is provided an implementation of the N weighted round robin by a $N.(2^W-1)$ way unweighted round robin where W is the number of bits defining a weight using a list constructed by interleaving N words of (2^W-1) bits each, with w_n 1's in a word where w_n is the weight of the queue n.

Figure 1 shows a hierarchical switch. The central interconnect ① provides the cross-interconnect between a number of high bandwidth ports. A set of multiplexers on the ingress side and demultiplexers on the egress side provide the aggregation function between the low and high bandwidth ports. The low bandwidth ports provide connections from the switch to the data sources on the ingress side and the data destinations on the egress side. In practice, a switch is required to support full-duplex ports, so an ingress multiplexer and its corresponding demultiplexer can be considered a single full-duplex device which will be termed here a *Router*.

It should be noted that the central interconnect ① may itself be a hierarchical switch, i.e., the methods described in this document can be applied to switches with an arbitrary number of hierarchical levels.

The aim of these methods is to provide a mechanism whereby the data stream from the switch to a particular destination, which comprises of a sequence of cells interleaved from various data sources, can be controlled such that predetermined proportions of its bandwidth are guaranteed to cells from each data source.

ddition to the data flow indicated by the arrows in figure 1 above, there is also a flow of backpressure or -control information associated with each of the data flows. This control-flow is indicated in figure 2 *dashed* arrows.

Bandwidth Allocation in the Ingress Multiplexer

An ingress multiplexer receives a set of data streams from the data sources via a set of low bandwidth in-ports. Each data stream is a sequence of equal size cells (equal number of bits of data). Figure 2 shows the architecture of the ingress multiplexer.

A set of N low bandwidth ports ① each fill one of N input queues . An Ingress Control Unit extracts the destination addresses from the cells in the input queues and transfers them into a set of M virtual output queues. There is one virtual output queue for each low-bandwidth output port in the switch.
The ingress multiplexer contains an NxM entry Ingress Port Table (IPT) which defines the how its bandwidth to a particular egress ports (via a particular virtual output queue) is distributed across the input ports. This table is used by the Ingress Control Unit to determine when (and to what degree) to exert backpressure to the data source resolved down to an individual virtual output queue.

The ingress multiplexer sends control information to the central interconnect indicating the state of the virtual output queues (*connection requests*). The central interconnect responds with a sequence of connections which it will establish between the routers (*connection grants*). The ingress multiplexer must now allocate the bandwidth to each egress demultiplexer provided by the central interconnect across the virtual output queues associated with each egress demultiplexer. The ingress multiplexer contains an Interconnect Link Control Unit ILCU which implements this function by scheduling cells from the virtual output queues across the high bandwidth link to the central interconnect according to an M entry Egress Port Table (EPT) .

Weighted Round Robin Implementation

The deterministic scheduling function of the ILCU can be defined as a weighted round robin arbiter (*WRR*). The ILCU receives a connection grant to a particular egress demultiplexer from the central interconnect and must select one of the N virtual output queues associated with that egress demultiplexer. This can be implemented by expanding the N way WRR into a $(N \cdot (2^W-1))$ way *unweighted* round robin, where W=(number of bits to define weight) such that is a queue has a weight of w, then it is represented as w-*1* entries in the unweighted round robin list (figure 3). The unweighted round robin list is constructed by interleaving N words (e_n) of (2^W-1) bits each, with w_n 1's in a word where w_n is the weight of the queue n as shown in Figure 3.

e.g., with 4 bit weights, a 4 way weighted round robin expands to a 60 way unweighted round robin.

In order to optimise the service intervals to the queues under all weighting conditions, the entries in the unweighted round robin list are distributed such that for each weight the entries are an equal number of steps apart (±1 step).

Table 1 shows an example of such an arrangement for 3 bit weights:

Table 1

w_n	e_n
1	1000000
2	1000100
3	1001010
4	1010101
5	1011011
6	1110111
7	1111111

In the Terachannel the arbiter must select one of 9 queues with 4 bit weights: 8 VOQs as described above and a multicast queue. This expands to a 135 entry unweighted round robin. The implementation of a large unweighted round robin arbiter can be achieved without resorting to a slow iterative shift-and-test method by "divide and conquer" – the 135 entry round robin is segmented into 9 off 16 entry round robins (as shown in figure 4) each of which can be implemented efficiently with combinatorial logic (9x16 provides upto 144 entries, so the multicast queue (upto 24 entries) actually can be allocated more bandwidth than an individual unicast queue (upto 15 entries)).

Figure 4 illustrates the partitioning of the round robin arbiter.

The *sorter* ① separates the request vector V (144 bits) into 9 off 16 bit vectors (v0-8). It also creates the 9 off pointers (p0-8) for each of the 16 bit round robin blocks . The block which corresponds to the existing pointer (which has been saved in register) gets a '1' at the corresponding bit position, while the other blocks get a dummy pointer initialised to location zero.
Each 16 bit round robin block now find the next '1' in its input vector and output its location (*g*), whether it had to wrap round (*w*) and whether it found a '1' in its vector (*f*). A *selector* can now identify the block which has found the '1' corresponding to the next '1' in the original 135 bit vector given a signal (*s*) from the sorter which specifies which round robin block had the original pointer position. The selector itself is a round robin function which can be implemented as combinatorial logic:

Find the next block starting at *s* which has *w* = false and *f* = true (if not found, select *s*).

Figure 5 shows an example of the above process, but for a smaller configuration for clarity,
(V = 12 bits, P = 4 bits, v0-2 = 4 bits, p0-2 = 2 bits, g0-2 = 2 bits).

Bandwidth Allocation in the Central Interconnect

The central interconnect provides the cross-connect function in the switch. The bandwidth allocation in the central interconnect is defined by an $(M/N)^2$ entry Central Allocation Table (CAT) which specifies the weights allocated to each possible connection through the central interconnect (the central interconnect has M/N high bandwidth ports).

A technique for bandwidth allocation in the central interconnect is described in another patent application ***Probabilistic Masking for Bandwidth Allocation.*** *(not yet filed)*

Bandwidth Allocation Tables (IPT, EPT, CAT) Programming
- what goes into the tables?

The Central Allocation Table (CAT) contains P^2 entries, where P=(M/N). Each entry w_{ie} defines the weight allocated to the connection from high bandwidth port *i* to high bandwidth port *e*. However, not all combinations of entries constitute a self consistent set, i.e., the allocations as seen from the outputs could contradict the allocations as seen from the inputs. A set of allocations is only self consistent if the sums of

weights at each output and input are equal. Figure 6 shows a self consistent and a non self consistent set of locations for a 4 port interconnect with 3 bit weights.

Assuming that the CAT has a self consistent set of entries, it is possible to define the bandwidth allocation to a link between input port i and output port e with weight w_{ie} as p_{ie}:

$$p_{ie} = \frac{w_{ie}}{\left(\sum_{n=0}^{(P-1)} w_{in}\right)}$$

The Egress Port Table (EPT) defines how the bandwidth of a high bandwidth port to the central interconnect is allocated across the virtual output queues. There is no issue with self consistency (all possible entries are self consistent), so the bandwidth allocation for a virtual output queue v with weight w_v is given by:

$$p_v = \frac{w_v}{\left(\sum_{n=0}^{(N-1)} w_n\right)}$$

Similarly, the Ingress Port Table (IPT) entries allocation the bandwidth of a virtual output queue to the ingress ports with port f with weight w_f given:

$$p_f = \frac{w_f}{\left(\sum_{n=0}^{(N-1)} w_n\right)}$$

Therefore the proportion of bandwidth at an egress port v allocated to an ingress port f is given by:

$$p_{fv} = p_f \cdot p_v \cdot p_{ie}$$

Managing Bandwidth Allocation Tables

– where do the weights come from?

In a switch which is required to maintain strict bandwidth allocation between ports (such as an ATM switch), the tables are setup via a Switch Management Interface (SMI) from a Connection Admission and Control (CAC) processor. When the CAC has checked that it has the resources available in the switch to satisfy the connection request, it can modify the IPT, EPT and CAT to reflect the new distribution of traffic through the switch.

In contrast, a switch may be required to provide a "best effort" service. In this case, the table entries are derived from a number of *local* parameters. Two such parameters are l_v (*length of virtual output queue v*) and u_v (*urgency of virtual output queue v*). Queue urgency is a parameter which is derived from the headers of the cells entering the queue from the ingress ports

A switch can be implemented which can satisfy a range of requirements (including the two above) by defining a Weighting Function which "mixes" a number of scheduling parameters to generate the table entries in real time according to a set of *sensitivities* to length, urgency and pseudo-static bandwidth

allocation (s_l, s_u, s_s). The requirement on the function are that it should be fast and efficient since multiple instances occur in the critical path of a switch. In the Terachannel the weighting function has the form:

$$w_v = \left(\frac{l_v^2}{2^{(1/sl)}} + \frac{p_v}{2^{(1/ss)}} + \frac{u_v}{2^{(1/su)}} \right) \cdot (1 - b_v)$$

Where b_v is the backpressure applied from the egress Router,
w_v is the weight of the queue as applied to the scheduler,
p_v is a pseudo static bandwidth allocation (e.g., EPT entry).

Backpressure in the Terachannel is described in another patent application ***Flow Control Architecture for a Digital Traffic Switch*** *(not yet submitted)*

Despite the apparent complexity of this function, it can be implemented exclusively with an adder, multiplexers and small lookup tables, thus meeting the requirement for speed and efficiency.

Features of this weighting function:

- s_l=1.0, s_s=0.0, s_u=0.0 : Bandwidth is allocated locally purely on the basis of queue length, with a non-linear transfer function, so that the switch always attempts to avoid queues overflowing.
- s_l=0.0, s_s=1.0, s_u=0.0 : Bandwidth is allocated purely on the basis of pseudo-static allocation as described above.
- s_l=0.0, s_s=1.0, s_u=0.5 : Bandwidth is allocated on the basis of pseudo-static allocation, but a data source is allowed to "push" some data flows harder (when the demand arises) by setting the urgency bits in the appropriate cell headers.

ample

ure 7 is a block diagram of a small switch based on the above principles, showing the correct numbers of queues, table and table entries.

e.g., for a hierarchical switch with 2 routers, each with 2 low-bandwidth ports (N=2, M=4),

Assuming that each low-bandwidth port can transport 1Gbps of traffic, each high-bandwidth link can carry 2Gbps and the switch is required to guarantee the following bandwidths allocations:

Flow bandwidth (Gbps)		Destination Port			
		A	**B**	**C**	**D**
Source Port	**A**	0.5	0.1	0.1	0.2
	B	0.2	0.2	0.2	0.2
	C	-	0.5	-	0.2
	D	0.1	0.1	0.6	0.2

The IPT, EPT and CAT tables would be set up by the CAC processor with the following 4 bit values (note that there will be rounding errors due to the limited resolution of the 4 bit weights):

IPT
(in Router AB):

	Source	
	A	**B**
A	15	6
B	3	6
C	3	6
D	6	6

(in Router CD):

	Source	
	C	**D**
A	0	3
B	15	3
C	0	15
D	6	5

EPT
(in Router AB):

	Source
	AB
A	15
B	6
C	6
D	8

(in Router CD):

	Source
	CD
A	2
B	12
C	12
D	12

CAT:

		Destination Router	
		AB	**CB**
	AB	15	10
	CD	10	15

Figures

Figure 1

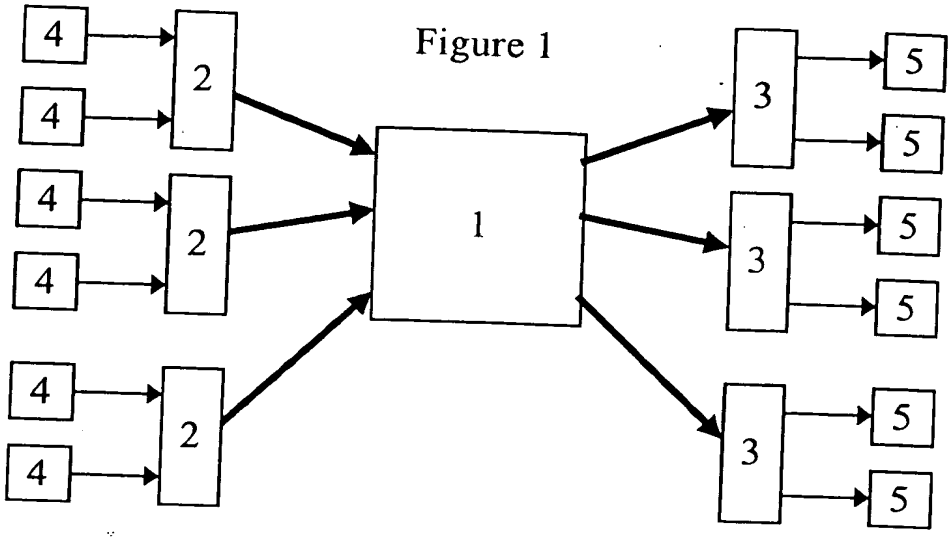

Figure 2

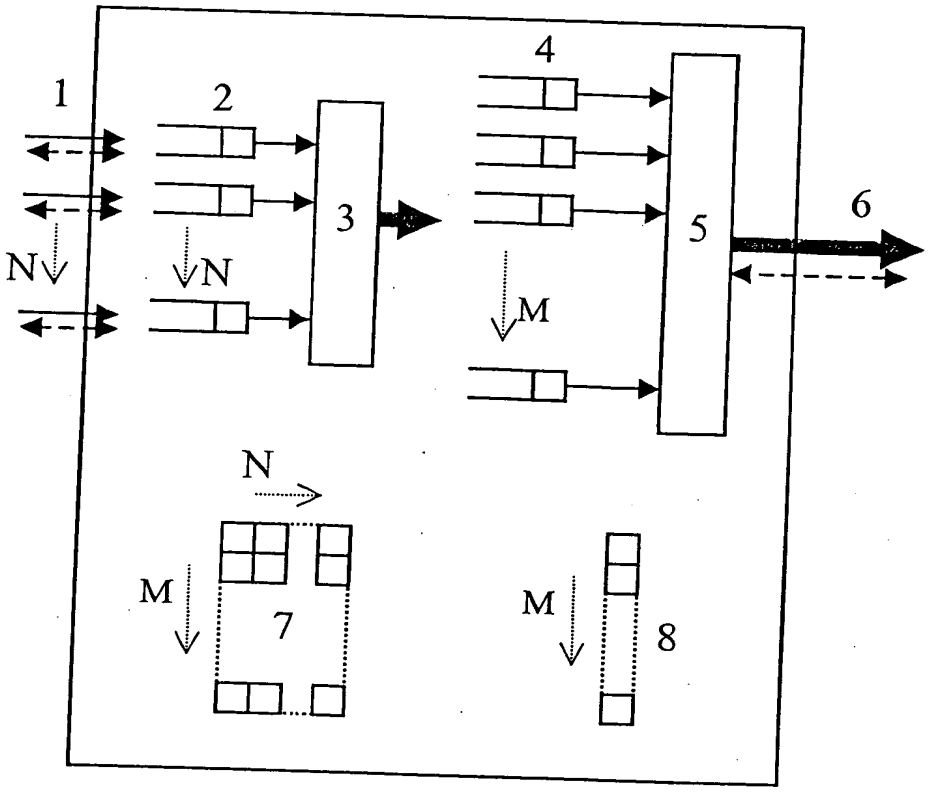

Figure 3

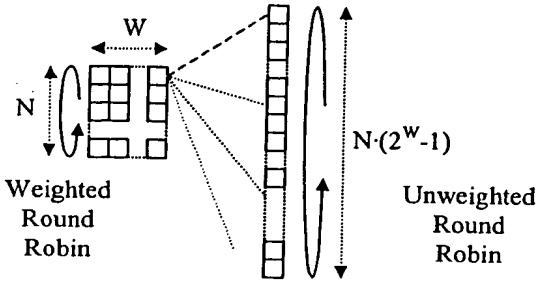

Figure 4

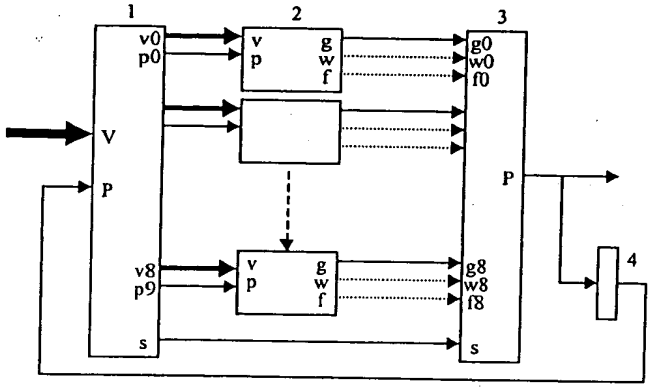

Figure 5